Prognostic breast cancer signature identified from 3D culture model accurately predicts clinical outcome across independent datasets

Katherine J. Martin[1], Denis R. Patrick[2], Mina J. Bissell[3] and Marcia V. Fournier[2 *]

[1] Bioarray Consulting, Belmont MA USA

[2] Department of Oncology-Biology, Oncology Center of Excellence for Drug Discovery, GlaxoSmithKline. Collegeville PA USA

[3] Cancer Biology, Lawrence Berkeley National Laboratory, Berkeley CA USA

*Corresponding author:
Marcia V. Fournier
Department of Oncology-Biology, Oncology CEDD, GlaxoSmithKline
1250 S Collegeville Rd UP 1450
Collegeville PA 19426 USA
E-mail: Marcia.V.Fournier@GSK.com

Running title: Prognostic role of 3D-signature in breast cancer.
Key words: 3D-culture, prognostic signature, breast cancer, microarray, estrogen receptor.

Abstract

One of the major tenets in breast cancer research is that early detection is vital for patient survival by increasing treatment options. To that end, we have previously used a novel unsupervised approach to identify a set of genes that predict prognosis of breast cancer patients. The predictive genes were selected in a well-defined cell culture model of non-malignant human mammary epithelial cell morphogenesis. Predictive genes were down-regulated during breast epithelial cell acinar formation and cell cycle arrest, using a three dimensional (3D) tissue culture in laminin-rich extracellular matrix. Here we examine the ability of this gene-signature (3D-signature) to predict prognosis in two independent large breast cancer microarray datasets having 286 samples in one and 122 samples in the other. Our results show that the 3D signature accurately predicts prognosis in two unrelated patient datasets. We also examine the correlation of individual genes with clinical outcome and estrogen receptor (ER) status. Poor prognosis tumors with a shorter time to relapse were generally associated with a higher level of expression of individual genes. The 3D-signature holds prognostic value for both ER-positive and ER-negative breast cancer and includes genes related to cell cycle, cell proliferation, angiogenesis and motility.

Introduction

Breast cancer ranks as the second leading cause of death among women with cancer in the US. Early detection of breast cancer has a significant impact on patient survival, though a portion of patients still relapse and rapidly develop a more aggressive form of disease (1). The identification of individuals with a high risk of relapse has become a primary focus of cancer research. Key steps are determining which patients will benefit from standard care therapies and assessing their chances of disease progression. Accurate identification of high-risk genes may not only lead to the identification of groups of high-risk patients, but also to the discovery of novel therapeutic molecular targets.

Several large studies have been performed to identify predictive signatures in breast cancer (2). These signatures have been selected using supervised methods applied to training sets of about 50-100 patients, and then confirmed in larger related sets ranging from 100-300 patients. It has been observed that the individual genes that comprise signatures identified in different studies show surprisingly little overlap. Investigations addressing this lack of overlap, have found that predictive signatures are highly dependent on the specific set of patients that make up the training set (3). Such disparity in signatures is not limited to breast cancer, but also has been found in schizophrenia studies. Less well studied is whether a given predictive signature that has been identified using a given dataset is also predictive in additional unrelated datasets.

Two predictive signatures for breast cancer identified by microarray analysis have been further developed into clinical multi-gene panel tests (4). MammaPrint became the first test approved by FDA for predicting breast cancer relapse and is composed of 70 genes. Oncotype DX, a prognostic test for ER positive breast cancers, has been commercially available since 2004 and is composed of 21 genes. The 70-gene signature was identified by analyzing at the large NKI dataset of van de Vijver et al. Unfortunately, subsequent analysis found that this signature did not predict outcome as well in an independent dataset (5). Several clinical trials are ongoing to test the utility of these prognostic gene-signature tests (6).

Even though gene signatures so far have been helpful for identifying patients at risk, they provide limited information on which genes are relevant to breast cancer biology. It follows that all genes included in gene-signatures cannot be key biological players in cancer progression. We hypothesize that the ability of a signature to demonstrate predictive power across different independent datasets tends to support the conclusion that it is composed of key, biologically relevant genes. The development of novel, biologically-based gene selection approaches may help to find these genes. We applied an unsupervised approach that is not dependent on the composition of a training set. The approach is based on a well-studied and biologically relevant model system that mimics cellular characteristics of human mammary gland. Since the genes are selected based on a biological parameter, they hold promise to represent key biological processes of cancer.

To select a prognosis signature for breast cancer, we used a 3D culture model of non-malignant human mammary epithelial cells (HMEC) (7). Non-malignant HMEC

reacquire the ability to form acini-like structures presenting a hollow lumen, basal polarity and cell cycle arrest in laminin-rich extracellular matrix 3D culture model. These acini structure recapitulate many of the characteristics of luminal cell differentiation in the mammary gland (8, 9). Here we describe the predictive power of a small set of 22 genes that were down-regulated during growth arrest and acini formation of HMEC in 3D cultures (3D-signature) in two large, independent breast cancer microarray datasets.

Materials and Methods

Dataset sources. The Wang dataset, consisting of the microarray profiles of 286 human breast tumors with associated clinical data (10), was obtained from GEO (Series GSE2034). The downloaded data were transformed to set measurements less than 25 to 25, chips and genes were median normalized and median polished. The Sorlie dataset, with microarray profiles of 122 human breast tumors and associated clinical data (11), was obtained from GEO (Series 4335). Downloaded data were transformed from log base 2 to linear values, then chips and genes were median-normalized. The van de Vijver dataset, with profiles of 295 human breast tumors and associated clinical data (12), was obtained from Rosetta Inpharmatics (http://www.rii.com/publications/2002/nejm.html). Downloaded data were transformed from log base 2 to linear values; chips and genes were median-normalized. Data processing steps were performed using GeneSpring software.

Kaplan-Meier curves: For survival analysis with the 3D-signature, patients were stratified into two groups using GeneSpring software by hierarchical cluster analysis and the expression levels of all 22 genes. Kaplan-Meier survival curves and log-rank statistics for these two groups were computed using MedCalc software. For survival analysis of the 22 individual signature genes, patients were stratified into quartiles for expression of each marker and survival curves were computed. Statistical analyses were performed using MedCalc software.

Results

We have previously used a novel unsupervised approach to identify a set of 22 genes that predict prognosis of breast cancer patients (7). This signature included genes that were down-regulated during breast epithelial cell acinar formation in 3D cultures in laminin-rich extracellular matrix (3D lrECM). Identities, Affymetrix IDs, GeneBank accession numbers, and biological functions of these genes are tabulated (**Supplemental data table 1**).

To further assess the utility of this 3D-signature, we have used two large independent breast cancer microarray datasets, both of which include annotated microarray data and associated clinical information. The dataset of Wang *et al*, includes data from 286 breast cancer patients while that of Sorlie, *et al*. (Stanford/Norway) includes data from 122 patients. Together these datasets represent a total of 408 patients. Numerous differences

exist between the datasets. Most notably, the patients were selected by different institutions using different admittance criteria. Database criteria are tabulated (**Supplemental data table 2**). We have previously shown that the 3D-signature predicted prognosis by using the dataset of van de Vijver, *et al*. (7). We used unsupervised hierarchical cluster analysis to group tumors into classes according to their gene expression patterns. Here we apply the same approach to the test the 3D-signature's ability to predict prognosis in the two additional datasets.

Probes for all 22 genes were present on the Affymetrix HG-U133A microarrays used by Wang, *et al*. Hierarchical cluster analysis found that the gene expression patterns separated the patients in two approximately equally sized clusters. Kaplan-Meier analysis was performed using relapse as an endpoint. The two clusters were highly significantly associated with prognosis (p=0.000013, Kaplan-Meier) (**Figure 1 A, B**).

Using the same approach, we also tested the dataset of Sorlie, *et al*. This dataset used Stanford two-color spotted microarrays of varying formats. Data for at least 40% of patients was available for 15 of the 22 3D-signature genes. Again, we applied hierarchical cluster analysis and found that the tumors were grouped in two approximately equally sized clusters. Kaplan-Meier analysis was performed using patient death as an endpoint. Results showed that the two clusters were significantly associated with prognosis (p=0.045, Kaplan-Meier) (**Figure 1 C, D**).

The ability of each individual gene of the 3D-signature to predict survival or relapse was tested by Kaplan-Meier analysis. For the Wang dataset, the expression levels of nine genes were significant predictors of a patient's time to relapse (**Figure 2A**). These genes included ASPM, AURKA, ACTN1, CEP55, CKS2, DUSP4, EPHA2, TRIP13, and VRK1. For each of these genes except DUSP4, poor prognosis tumors with a short time to relapse were associated with a higher level of gene expression (>2-fold increase). For DUSP4, the pattern was reversed and poor prognosis was associated with a lower level of expression (> 2 fold decrease). For the Sorlie dataset, expression levels of seven genes were significant predictors of survival time (**Figure 2B**). These genes included AURKA, CDKN3, CEP55, FOXM1, RRM2, TRIP13, and VRK1. For all of these genes, poor prognosis was associated with a higher level of gene expression. These Kaplan-Meier p-values are summarized in **Table 1**, which also lists our previously determined p-values from the van de Vijver dataset for comparison. The results show that 41% (9 of 22), 39% (7 of 18), and 68% (13 of 19) of the genes were significant individual predictors in the Wang, Sorlie, and van de Vijver datasets, respectively **(Table 1)**.

Table 1 groups the 3D-signature genes by the biological process in which they participate. The genes include five categories: cell cycle/mitosis, motility/angiogenesis, polyamine biosynthesis, and transcription/replication genes and one gene of unknown function.

We have also looked at the ability of the genes to predict prognosis in ER+ and ER- subsets of patients. **Table 1** lists Kaplan-Meier p-values for ER+ and ER- tumors for all three datasets. In the Wang dataset, more of the 3D-signature genes tended to associate

with prognosis of ER+ tumors ($p<0.1$) than ER- tumors (Fishers exact test, $p=0.034$), though in the other two datasets, there was no statistical difference in the numbers of markers for ER+ and ER- tumors among the 22 genes (**Table 1**).

A notable finding among the ER related differences was that the genes that tended to associate with prognosis in ER+ patients had different molecular functions than the genes that tended to associate with prognosis in ER- patients. In particular, significantly more cell cycle and transcription genes were prognostic markers for ER+ tumors (Fisher's exact test, $p= 0.0047$), while prognostic markers of ER- tumors were significantly more likely to have functions related to angiogenesis and metastasis (Fisher's exact test, $p=0.023$). This analysis considered results from all three of the datasets.

To summarize, the genes that tended to associate with prognosis in ER+ tumors ($p<0.1$ for at least one of the three datasets) included AURKA, CDKN3, CEP55, DUSP4, NCAPG, RRM2, ACTB, EPHA2, FGFBP1, TNFRSF6B, EIF4A1, and VRK1 (**Table 1**). Genes that tended to associate with prognosis in ER- tumors included TUBG1, ACTB, FGFBP1, FOXM1, SERPINE2, and TNFRSF6B. Genes that were markers for prognosis in both ER+ and ER- tumors included ACTB, FGFBP1, and TNFRSF6B.

We have also tested for an association between expression of the individual 3D-signature genes and tumor ER status. (**Table 2**). We note that expression levels of the majority of the 22 genes were significantly associated with ER status. For the Wang, Sorlie, and van de Vijver patient datasets, percentages associated with ER status were 91%, 71%, and 84%, respectively. There was a very strong statistical enrichment for ER status related genes among the 3D-signature genes (Fisher's exact test, $p= 3.11e-8$, Wang dataset). In the Wang dataset, the expression levels of 20 of the 22 signature genes (91%) were significantly associated with ER status, while, for the entire set of 22,283 genes, expression levels of a total of 7,424 genes (33%) were ER associated (Welch t-test with FDR, $p<0.05$).

The genes that correlated with ER status also correlated with basal/luminal status (Fisher's exact test $p=0.011$) (data not shown). The majority of the genes were more highly expressed in ER- breast cancers than ER+ breast cancers. Two genes (DUSP4 and TUBG1) had the reverse pattern and were significantly under-expressed in ER-negative tumors (correlation analysis, $p<0.05$). In the Wang dataset, we found that the highly ER-associated genes were no more likely to be good prognostic markers than the more poorly ER-associated genes (Fishers exact tests, $p<0.05$) (**Table 1**). This conclusion applied to the subsets of ER+ tumors and ER- tumors, as well as all patients.

Discussion

We hypothesized that the changes in gene expression occurring during acini formation of non-malignant HMEC in a 3D culture model are opposite from those occurring during the development of breast tumors with a poor prognosis. In support of this hypothesis, we showed that genes that were expressed at significantly lower levels in organized, growth

arrested HMEC than in their proliferating counterparts could be used to classify breast cancer patients into poor and good prognosis groups (7). The present study provides two independent confirmations of a 22 gene prognostic signature (3D-signature) that we previously identified using a novel unsupervised strategy.

One of the key criticisms of gene signatures identified using microarray technology is the lack of validation across platforms (11, 13). Here we report that the 3D-signature predicted prognosis in two large independent datasets ($p=0.00001$ and 0.045 for datasets of Wang et al and Sorlie et al, respectively; Kaplan-Meier analysis). To date, the 3D-signature has been tested in three large datasets for a total of 703 breast cancer patients. There were differences in how well the signature performed between the datasets. Prognosis was best for the Wang dataset. Microarrays used for this dataset were identical to those of our selection study and included probes for all 22 genes. In contrast, microarrays used for the Sorlie dataset included probes for only 15 of the 22 genes, and some of these 15 probes could potentially recognize different isoforms of the genes than those of the selection study. However, even with these differences with probe composition, the 3D-signature accurately predicted prognosis in both datasets.

The 3D-signature includes cell cycle and transcription related genes that predict prognosis in ER+ breast cancer patients. This finding is consistent with previous studies that show that proliferation and cell cycle genes are the strongest predictor for relapse among ER positive patients (14). In several previous studies, a signature enriched in cell cycle related genes has been reported to predict poor prognosis of breast cancer, along with a second smaller class of genes that includes transcription related genes. Poor prognosis in ER+ tumors in particular has been found to be strongly predicted by over expression of cell cycle and cell proliferation genes (10, 15-17) .

The 3D-signature also includes angiogenesis and motility genes that are markers for prognosis in both ER+ and ER- tumors. Genes in this functional class of breast tumor marker genes were also identified in other breast cancer signatures (15, 18), though the association of this functional class with ER- tumors has not been noted for gene signatures. Markers for ER- tumors have been reported to be significantly less prevalent than markers for ER+ tumors (17). Some genes within this functional class predicted prognosis for only ER+ tumors, some predicted prognosis for only ER- tumors, and some predicted prognosis for both ER+ and ER- tumors.

Since few overlaps have been found among the published breast cancer signatures, it appears that many (thousands) of marker genes have predictive ability in different subsets of patients. It has been proposed that some genes may have moderate predictive ability in many patients, while some may be “master genes” with high predictive ability in as yet undefined subsets of patients. When many such genes are used together, a highly accurate predictive tool results that is accurate across a wide cross section of breast cancer patients. The actual composition of the signature may be less important than the fact the each signature is a set of many semi-predictive genes. In contrast to gene signatures identified from specific patient sets by supervised methods, our approach is based on a well studied and biologically relevant model system that mimics the human mammary

gland. Hence the 3D-signature holds promise to include "master genes" of key biological processes of cancer.

Earlier detection can benefit patient survival and treatment options; however progress is still needed in developing therapeutic strategies amenable to early stage disease. A focus on the development of novel treatments targeting early disease rather than advanced malignant carcinoma seems to be a natural next step. The identification of key regulatory pathways that maintain the self-limited proliferation of non-malignant cells in 3D cultures may direct us to novel molecular targets for earlier cancer therapy.

Acknowledgements
We thank Mary Ann Hardwicke for critical reading of the manuscript and for helpful comments. This work was also supported through LBNL Contract No. DE-AC02-05CH11231.

Figure Legends

Figure 1. The 22 gene 3D signature predicts survival in the microarray datasets of Wang, *et al*., and Sorlie, *et al*. The 22 gene signature and unsupervised hierarchical clustering grouped breast cancer patients to accurately reflect overall relapse or survival when analyzed by the method of Kaplan and Meier. **A.** Hierarchical cluster analysis of the dataset of Wang, *et al*. The pattern of expression of the 22 genes selected by the 3D assay are shown for the 286 breast cancer patients of Wang et al. Genes and samples were organized by using hierarchical clustering. The two major clusters in the sample dimension (red cluster and yellow cluster), were found by using survival analysis to distinguish between good and poor prognosis patients ($p < 0.0001$). **B.** Kaplan-Meier curves for the red and yellow clusters of the hierarchical diagram of panel A. The endpoint recorded for this dataset was relapse, measured in months. **C.** Hierarchical cluster analysis of Sorlie, *et al*. dataset. The pattern of expression of the 15 of 22 genes with probes on the Stanford microarrays and with data available for at least 40% of patients are shown for the 121 breast cancer patients reported by Sorlie *et al*. Expression was organized by hierarchical clustering. The two major clusters in the sample dimension (red cluster and yellow cluster), were found by using survival analysis to distinguish between good and poor prognosis patients ($p = 0.00447$). **D.** Kaplan-Meier curves for the red and yellow clusters of the hierarchical diagram of panel C. The endpoint recorded for this dataset was death, measured in months.

Figure 2. Kaplan-Meier curves of the individual genes that accurately predicted patient prognosis ($p<0.05$). A. Results for individual genes in the dataset of Wang, et al. using patient relapse as the endpoint. **B.** Results for individual genes in the dataset of Sorlie, et al. using patient survival as the endpoint.

Supplemental Figure 1. Kaplan-Meier curves of the individual genes that did not accurately predict patient prognosis ($p>0.05$). A. Results for individual genes in the dataset of Wang, et al. using patient relapse as the endpoint. **B.** Results for individual genes in the dataset of Sorlie, et al. using patient survival as the endpoint.

References

1. Cancer Facts & Figures. *In:* AC Society (ed.). Atlanta: American Cancer Society, 2007.
2. Edgren, H and Kallioniemi, O Integrated breast cancer genomics. Cancer Cell 2006; *10:* 453-4.
3. Ein-Dor, L, Zuk, O, and Domany, E Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci U S A 2006; *103:* 5923-8.
4. Hinestrosa, MC, Dickersin, K, Klein, P, et al. Shaping the future of biomarker research in breast cancer to ensure clinical relevance. Nat Rev Cancer 2007; *7:* 309-15.
5. Gruvberger, SK, Ringner, M, Eden, P, et al. Expression profiling to predict outcome in breast cancer: the influence of sample selection. Breast Cancer Res 2003; *5:* 23-6.
6. Branca, M Genetics and medicine. Putting gene arrays to the test. Science 2003; *300:* 238.
7. Fournier, MV, Martin, KJ, Kenny, PA, et al. Gene expression signature in organized and growth-arrested mammary acini predicts good outcome in breast cancer. Cancer Res 2006; *66:* 7095-102.
8. Petersen, OW, Ronnov-Jessen, L, Howlett, AR, and Bissell, MJ Interaction with basement membrane serves to rapidly distinguish growth and differentiation pattern of normal and malignant human breast epithelial cells. Proc Natl Acad Sci U S A 1992; *89:* 9064-8.
9. Bissell, MJ Modelling molecular mechanisms of breast cancer and invasion: lessons from the normal gland. Biochem Soc Trans 2007; *35:* 18-22.
10. Wang, Y, Klijn, JG, Zhang, Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 2005; *365:* 671-9.
11. Sorlie, T, Tibshirani, R, Parker, J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A 2003; *100:* 8418-23.
12. van de Vijver, MJ, He, YD, van't Veer, LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002; *347:* 1999-2009.
13. Esteva, FJ and Hortobagyi, GN Prognostic molecular markers in early breast cancer. Breast Cancer Res 2004; *6:* 109-18.
14. Loi, S, Piccart, M, and Sotiriou, C The use of gene-expression profiling to better understand the clinical heterogeneity of estrogen receptor positive breast cancers and tamoxifen response. Crit Rev Oncol Hematol 2007; *61:* 187-94.
15. van 't Veer, LJ, Dai, H, van de Vijver, MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002; *415:* 530-6.
16. Sotiriou, C, Wirapati, P, Loi, S, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst 2006; *98:* 262-72.

17. Teschendorff, AE, Naderi, A, Barbosa-Morais, NL, et al. A consensus prognostic gene expression classifier for ER positive breast cancer. Genome Biol 2006; *7:* R101.
18. Chang, HY, Nuyten, DS, Sneddon, JB, et al. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. Proc Natl Acad Sci U S A 2005; *102:* 3738-43.

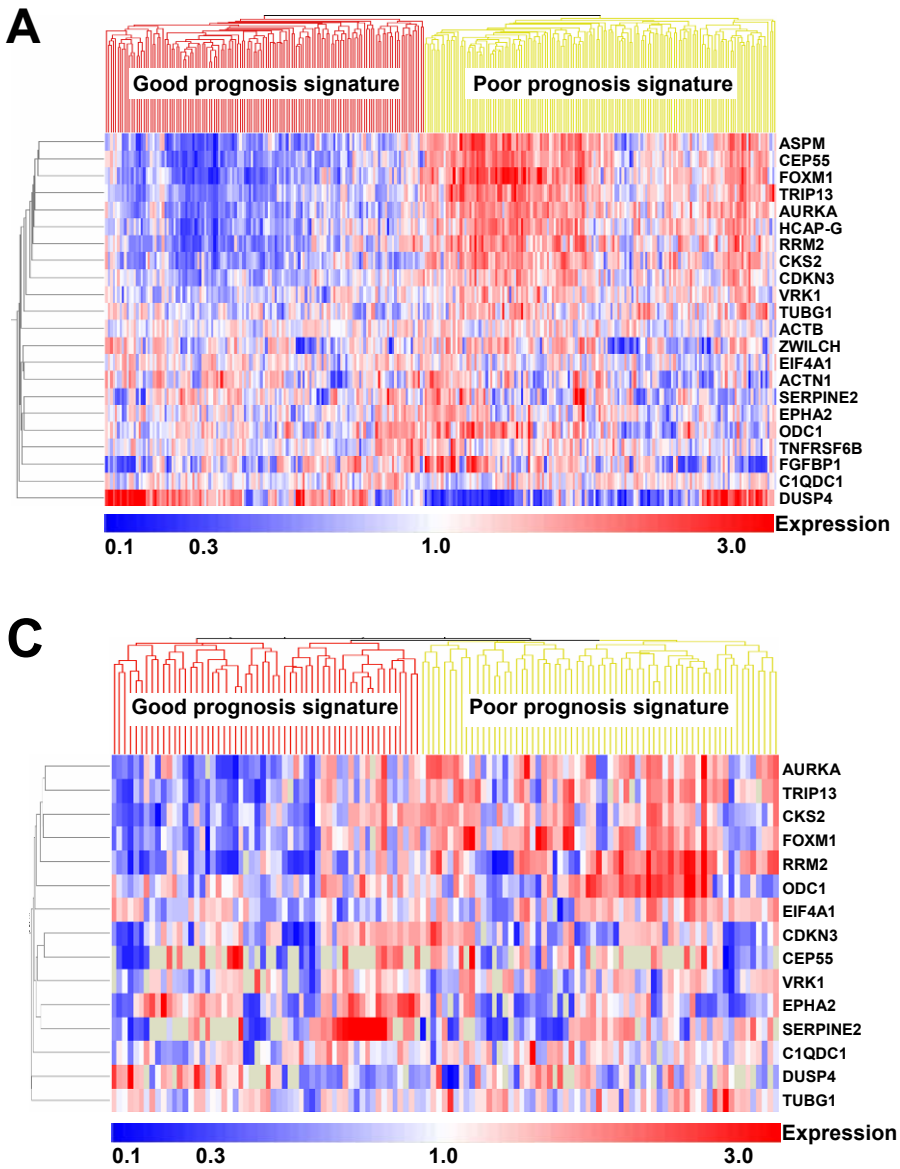
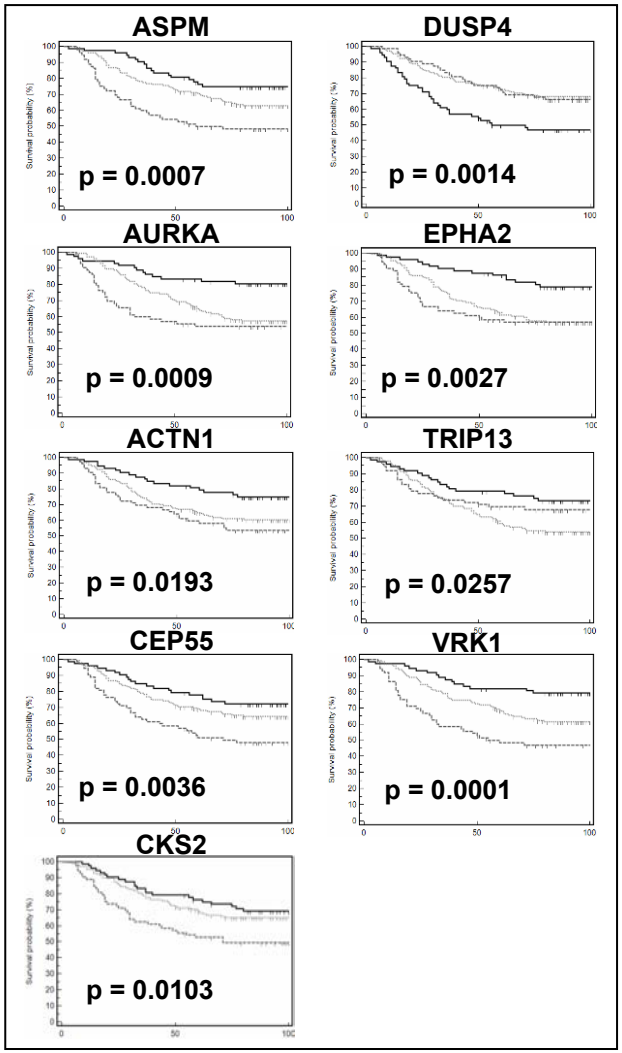

A
Good prognosis signature
Poor prognosis signature
ASPM
CEP55
FOXM1
TRIP13
AURKA
HCAP-G
RRM2
CKS2
CDKN3
VRK1
TUBG1
ACTB
ZWILCH
EIF4A1
ACTN1
SERPINE2
EPHA2
ODC1
TNFRSF6B
FGFBP1
C1QDC1
DUSP4
Expression
0.1
0.3
1.0
3.0
B
Wang et al. dataset
Relapse probability (%)
Red cluster
Yellow cluster
p = 0.000013
Time to relapse (months)
Number at risk
Group: red cluster
136 111 54 4
Group: yellow cluster
150 88 45 7
C
Good prognosis signature
Poor prognosis signature
AURKA
TRIP13
CKS2
FOXM1
RRM2
ODC1
EIF4A1
CDKN3
CEP55
VRK1
EPHA2
SERPINE2
C1QDC1
DUSP4
TUBG1
Expression
0.1
0.3
1.0
3.0
D
Sorlie et al. dataset
Survival probability (%)
Red cluster
Yellow cluster
p = 0.0447
Survival time (months)
Number at risk
Group: red cluster
65 8 0
Group: yellow cluster
56 20 5

Figure 1


A. Wang dataset
ASPM
p = 0.0007
DUSP4
p = 0.0014
AURKA
p = 0.0009
EPHA2
p = 0.0027
ACTN1
p = 0.0193
TRIP13
p = 0.0257
CEP55
p = 0.0036
VRK1
p = 0.0001
CKS2
p = 0.0103

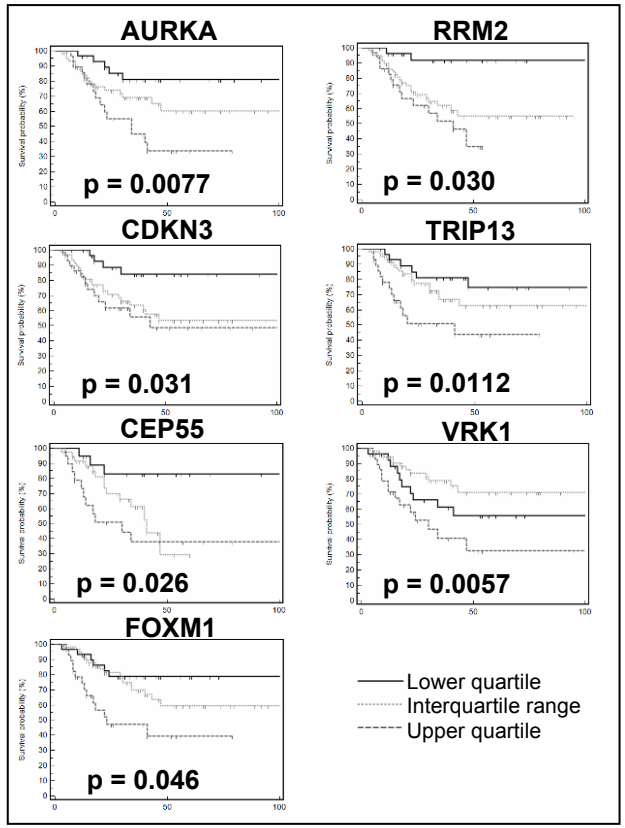

B. Solie dataset
AURKA
p = 0.0077
RRM2
p = 0.030
CDKN3
p = 0.031
TRIP13
p = 0.0112
CEP55
p = 0.026
VRK1
p = 0.0057
FOXM1
p = 0.046
Lower quartile
Interquartile range
Upper quartile

Figure 2

Table 1. Kaplan-Meier p-values for the 22 individual genes in the Wang, Sorlie, and van de Vijver patient datasets

	All patients			ER + patients			ER- patients		
Gene	**Wang**	**Sorlie**	**Vijver***	**Wang**	**Sorlie**	**Vijver**	**Wang**	**Sorlie**	**Vijver**
Cell cycle / mitosis genes									
ASPM	**0.0007**	-	ns	ns	-	ns	ns	-	ns
AURKA	**0.0009**	**0.0077**	**0.0001**	**0.0062**	0.067	ns	ns	ns	ns
CDKN3	ns	**0.031**	**0.0001**	ns	0.095	ns	ns	ns	ns
CEP55	**0.0036**	**0.026**	**0.0001**	ns	**0.023**	ns	ns	ns	ns
CKS2	**0.01**	ns	**0.05**	ns	ns	ns	ns	ns	ns
DUSP4	**0.0041**	ns	**0.0001**	0.085	ns	ns	ns	ns	ns
NCAPG	ns	-	-	0.087	-	-	ns	-	-
RRM2	ns	**0.03**	**0.0001**	ns	**0.028**	0.068	ns	ns	ns
TUBG1	ns	ns	**0.05**	ns	ns	ns	0.059	ns	ns
Angiogenesis / motility genes									
ACTB	ns	-	**0.05**	0.079	-	ns	ns	-	**0.041**
ACTN1	**0.019**	ns	**0.01**	ns	ns	ns	ns	ns	ns
EPHA2	**0.0027**	ns	**0.05**	**0.011**	ns	**0.026**	ns	ns	ns
FGFBP1	ns	ns	ns	ns	ns	**0.049**	ns	ns	0.059
FOXM1	ns	**0.046**	**0.0001**	ns	ns	ns	ns	ns	**0.039**
SERPINE2	ns	-	-	ns	ns	-	0.092	**0.0088**	-
TNFRSF6B	ns	ns	ns	0.076	ns	ns	ns	ns	0.071
ZWILCH	ns	ns	**0.01**	ns	ns	ns	ns	ns	ns
Polyamine biosynthesis									
ODC1	ns	ns	ns	ns	ns	ns	ns	ns	ns
Transcription / tranlsation genes									
EIF4A1	ns	ns	ns	0.095	ns	ns	ns	ns	ns
TRIP13	**0.0257**	**0.0112**	**0.0001**	ns	ns	ns	ns	ns	ns
VRK1	**0.0001**	**0.0057**	ns	**0.018**	**0.026**	ns	ns	ns	ns
Unknown function									
C1QDC1	ns	ns	-	ns	ns	ns	ns	ns	ns

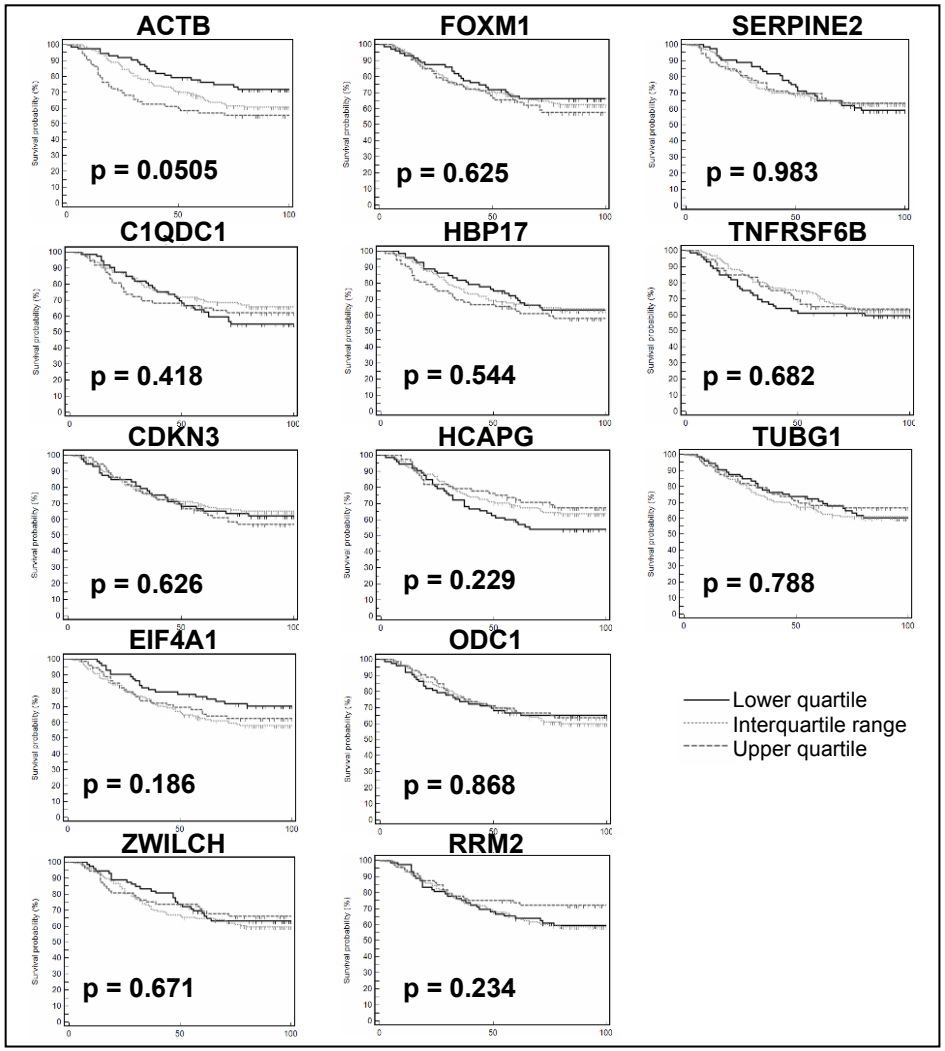
* Data previously reported (Fournier et al., Cancer Research 2007).
ns = not significant; - = no data; bold/underlined = p<0.05.

Table 2. ER association of the 22 individual genes in patient datasets from Wang, Sorlie, and van de Vijver (Welch t-test p values with false positive multigene correction).

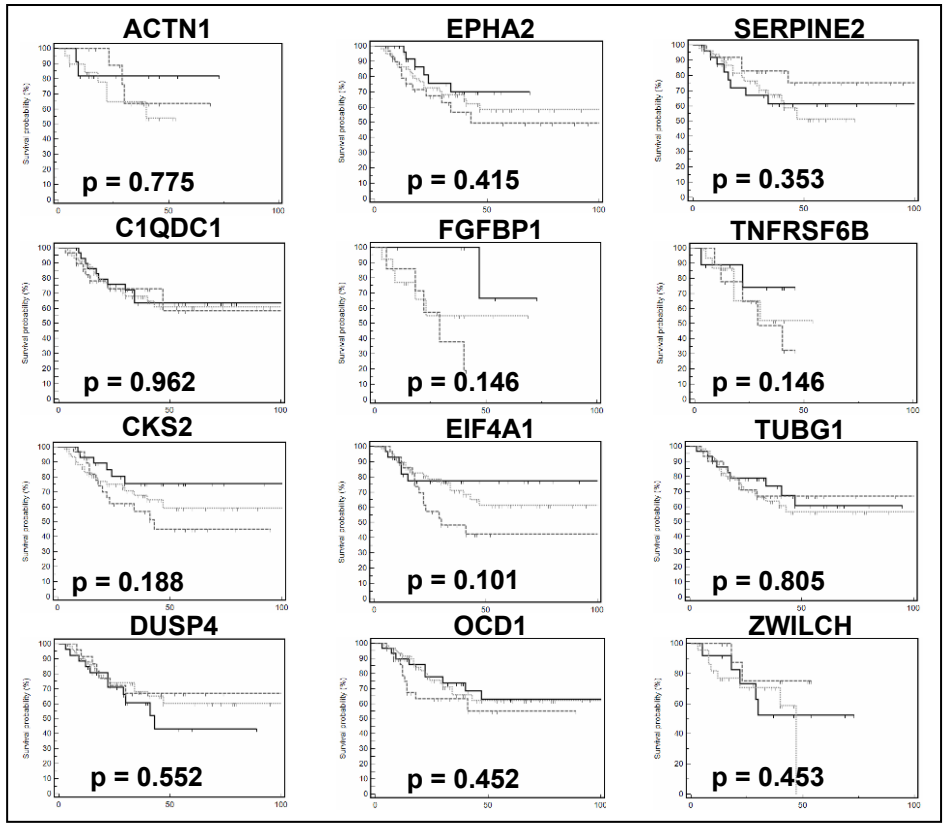
Gene	Wang	Sorlie	van de Vijver
Cell cycle / mitosis genes			
ASPM	3.1e-9	-	9.2e-5
AURKA	3.6e-8	0.018	1.2e-8
CDKN3	3.8e-6	ns	0.024
CEP55	4.5e-10	ns	6.9e-10
CKS2	0.0011	3.8e-9	0.0017
DUSP4	5.3e-7	0.044	6.1e-9
NCAPG	2.4e-5	-	-
RRM2	7.6e-12	0.049	2.7e-9
TUBG1	ns	0.018	ns
Angiogenesis / motility genes			
ACTB	0.018	-	5.8e-11
ACTN1	2.4e-5	ns	0.027
EPHA2	1.7e-9	0.028	1.2e-10
FGFBP1	1.6e-6	-	0.00069
FOXM1	1.7e-9	0.025	3.5e-9
SERPINE2	5.1e-5	ns	-
TNFRSF6B	0.0030	ns	0.0017
ZWILCH	0.0081	0.00018	0.0015
Polyamine biosynthesis			
ODC1	6.7e-11	ns	5.8e-11
Transcription / translation			
EIF4A1	0.018	0.018	ns
TRIP13	9.0e-9	0.0042	1.2e-10
VRK1	0.0061	4.2e-5	ns
Unknown function			
C1QDC1	ns	0.00094	-

ns = not significant; - = no data

A. Wang dataset genes with Kaplan p > 0.05



B. Sorlie dataset genes with Kaplan p > 0.05



Supplemental Figure 1

Supplemental Table 1. List of 22 genes in prognostic signature of Fournier et al. 2006

Symbol	Alias	Affymetrix ID	GenBank	Description	Functional class
ASPM	FLJ10517	219918_s_at	NM_018123	asp (abnormal spindle) homolog, microcephaly assoc	cell cycle/mitosis
AURKA	STK6	218542_at	NM_018131	aurora kinase A	cell cycle/mitosis
CDKN3	-	218456_at	NM_023925	cyclin-dep kinase inhib 3 (CDK2-assoc dual spec phos)	cell cycle/mitosis
CEP55	FLJ10540	209714_s_at	AF213033	centrosomal protein 55kDa	cell cycle/mitosis
CKS2	-	204170_s_at	NM_001827	CDC28 protein kinase regulatory subunit 2	cell cycle/mitosis
DUSP4	-	204014_at	NM_001394	dual specificity phosphatase 4	cell cycle/mitosis
NCAPG	HCAP-G	214805_at	U79273	chromosome condensation protein G	cell cycle/mitosis
RRM2	-	203499_at	NM_004431	ribonucleotide reductase M2 polypeptide	cell cycle/mitosis
TUBG1	-	205014_at	NM_005130	tubulin, gamma 1	cell cycle/mitosis
ACTB	-	202580_x_at	NM_021953	actin, beta	motility / angiogenesis
ACTN1	-	218663_at	NM_022346	actinin, alpha 1	motility / angiogenesis
EPHA2	-	209773_s_at	BC001886	EPH receptor A2	motility / angiogenesis
FGFBP1	HBP17	208637_x_at	BC003576	heparin-binding growth factor binding protein	motility / angiogenesis
FOXM1	-	203856_at	NM_003384	forkhead box M1	motility / angiogenesis
SERPINE2	-	212190_at	AL541302	serpin peptidase inhib 2 (nexin)	motility / angiogenesis
TNFRSF6B /RTEL1		200801_x_at	NM_001101	tumor necrosis factor receptor superfamily, 6b, decoy	motility / angiogenesis
ZWILCH	FLJ10036	218349_s_at	NM_017975	zwilch, kinetochore associated, homolog	motility / angiogenesis
ODC1	-	206467_x_at	NM_003823	ornithine decarboxylase 1	polyamine biosynthesis
EIF4A1	-	201714_at	NM_001070	eukaryotic translation initiation factor 4A, isoform 1	transcription/replication
TRIP13	-	204033_at	NM_004237	thyroid hormone receptor interactor 13	transcription/replication
VRK1	-	204092_s_at	NM_003600	vaccinia related kinase 1	transcription/replication
C1QDC1	CAPRIN2	200790_at	NM_002539	C1q domain containing 1	unknown function

Supplemental Table 2. Comparison of microarray datasets.

Publications		Microarrays			Patient samples		
First author	**Citation**	**Array platform**	**No. spot features**	**Signature genes on array**	**No. samples**	**Breast cancer patient population**	**End point**
van de Vijver	NEJM 2002; 347:1999-2009	Fluorescent array made by Rosetta Inpharmatics	25,000	19	295	Stage I or II invasive carcinoma patients less than 52 years of age	Death
Wang	Lancet 2005; 365:671–79	Affymetrix HG-U133A	23,000	22	286	Lymph-node negative patients with no systemic therapy	Relapse
Sorlie	PNAS 2003; 100:8418-23	Fluorescent arrays made at Stanford University	8 different platforms ranging from 9,200 to 54,000 features	10-19	122	Histology subtypes included 4 normal breast, 2 DCIS, 100 invasive ductal carcinoma, 3 fibroadenoma, 8 lobular carcinoma, and 1 each of mucinous, papillary, pleomorphic, and undifferentiated carcinomas	Death